INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Information Technology, Computing & Applications (ICITCA 2021)

Causal Discovery using Dimensionality Reduction Partial Association Tree

Sreeraman.Y[1], S. Lakshmana Pandian[2]
[1]Researh Scholar, Dept. of CSE, Pondicherry Engineering College, Puducherry, India.
[2]Associate Professor, Dept. of CSE, Pondicherry Engineering College, Puducherry, India.
sramany@gmail.com[1]

Abstract

Decision tree is a model to classify data based on labelled attribute values. This model is a supervised learning approach through which one can classify a new entry into an appropriate class. If we want to know the cause behind this classification then decision tree cannot provide the same. When we infer causes behind the classification then they will provide a rich knowledge for better decision making. Causal Bayesian Networks, Structural Equation Models, Potential Outcome Models are the some of the models that are used to get causal relationships. These models need experimental data. But it is not possible/ it is very expensive to conduct full experiments. So a model is needed to identify causes from effects from observational data rather than experimental data. In this paper a novel approach is proposed for causal inference rule mining which can infer the causes from observational data in a faster way and also scalable. Statistical tools and techniques named partial association test, correlation are used to develop the model. A new way of constructing a tree called Dimensionality Reduction Partial Association Tree (DRPAT) is introduced. Sometimes the existing causality cannot be extracted where low associated dimensions are involved in data and hiding the underlying causality and this model extracts causal association in case of hidden causality in data.. The model is applied on "Cardiovascular Disease dataset" sourced from Kaggle Progression System. The result is a Partial Association Tree. From this tree one can get a set of causal rules which can form a basis for better data analytics and then the better decision making.

Keywords: Decision tree, supervised learning, partial association tree, causal rules

1. Introduction

The relationship among data variables can be explored with the help of normal decision tree. It tests data context for a target variant or class label. Causal inference is playing a vital role in data analysis of big data sources particularly in medical and healthcare domains. Experimentation is needed to find causal relationships, but it is sometimes difficult to perform full experimentation due to cost, exhaustive nature of the experimental unit and time. Casual relationships always imply associations but the converse always may not be true. There is a large scope and need with respect to medical big data sources to mine causal relationships in data that can provide better means for decision making in medication and treatments. In this paper a causal relation framework in the form of a decision tree based on partial association in data is constructed with a couple of assumptions about association, partial association, decision purity and the decision strength at the concluding node .These assumptions leads to design option for causal decision tree. This framework can identify the casual relationship between the predictor variable(s) and the outcome variable. The next session is dedicated to review the attempts made in the literature to build causal inference models. The third session describes about the problems

identified through the review. The details of the proposed algorithm are given next. The description of the dataset followed by the results, discussion and conclusion come at the final sessions.

2. Literature Review

Extraction of causes from effects is a challenging task. This needs practical data but it is not possible to do the whole set of experiments as they need large amounts of time and efforts. Observational data provide the alternative to this problem. Association rule mining, correlations and the related statistical tests are providing means for finding causal indications. A statistical test named Mantel–Haenszel test [1] is used for repeated tests of independence when we have multiple 2x2 tables of independence. When there is an availability of observational data relating to a problem context, we can go for stratification of such data to convert it into multiple independent 2x2 tables. The causal relationship between the pair of variables is determined based on chi-square value of the test. In this way Mantel–Haenszel test provides us the tool to identify causality in data. Causal relationships can be found from the data that was collected on observations [2].Observational data needs hypothesis settings and these hypotheses undergone for testing using the observational or sample data. This type of study needs the expertise of domain experts. There is no perfect automation to do all these sequential tasks. If the automation is possible it must be competent with the ever increasing size of data. Though a decision tree is able to deal with observational data it is limited to the class prediction task [3]. A normal decision tree is not a suitable substitute for causal relationship mining. Therefore there is a need to construct a model that can identify the causality in data and can interpret the data context with respect to the underlying causalities. Causality is connected with probabilities and the probability needs to be quantified when deciding before a variable causes another variable [4]. But probabilities cannot eliminate confusion. Probabilistic causality played a good role in past research. Causal analysis is an upcoming area in medicine and social sciences. Conclusion of causal relationship between two variables needs a lot of care when the decision is used for clinical applications [5].Though these model choices are quite interesting in idea, the applicability of these models has their own limitations. Some are demanding high experimentation cost. Some are facing the scalability problems. For some contexts the research proposal set by the investigator regarding causal inference cannot be tested base on exhaustive nature of the research objects and as a result the experimental data is not possible. To deal with such situations observational study is a good alternative where the investigator does not intervene and rather simply "observes" and assesses the strength of the relationship between a predictor variable and a response variable[6].Recently some researchers proposing partial association rules to elevate casual relationships among data where the data is mostly observational [7][8].

3. Problem Description

Uncovering causality in data is always a challenging task. Generally full experimentation is needed for this task but it is not possible due to the cost and other limitations of the experiment. But Observational data play a vital role in cause identification, but set of assumptions and computational overloads are needed for the analysis of observational data for cause identification. The existing methods of causal discovery are suffering from scalability issues. The existing methods of causal relationship discovery have their own limitations in terms of experimentation cost, problem of high dimensionality, and scalability problems. Sometimes the nature of the data is also a significant limitation. To cope with such limitations better novel methodologies is demanding.

4. Proposed Model

The proposed model is a decision tree where each splitting attribute is decided based on the strength of causal relationship between the current attribute and the goal attribute. A statistical test named Mantel–Haenszel test for partial association is used to test the causality. The splitting attribute for the decision tree is the attribute with highest partial association value. The above process is continued at each node until the leaf nodes are generated based on some specified criteria.

4.1 Proposed Algorithm

Input

A = set of attributes (A1, A_2..., A_{k-1}, T) where k is the dimension of the dataset.

D = dataset

h = height of the tree
e = edge label (1 for left, 0 for right)

Output

Partial Association Tree (PAT)

Steps:

1. Create root node N
2. Compute the correlation of A_i versus T and trim the size of dimension by eliminating the attributes which are unable to meet the specified correlation threshold
3. Partial_Association_Tree_Generation (N, A, D, D', h, e)
4. Prune Partial Association Tree: Prune (N)

4.1.1 Partial_Association_Tree_Generation (N, A, D, D', h, e)

{
if (A is empty or (h + 1) = threshold) then
Create new treeNode and then find its class label based on majority labeling.
if (e = 1) Then
add treeNode as left child of parent node, N
else
add treeNode as right child of parent node, N
end if
return
end if
find Correlation of each Attribute in A
find PAT value for each correlation threshold satisfied attribute in A using the below formula

$$PAT(A_i, Y) = \frac{\left(\left|\sum_{k=1}^{r} \frac{n_{11k}n_{22k} - n_{21k}n_{12k}}{n_{..k}}\right| - \frac{1}{2}\right)^2}{\sum_{k=1}^{r} \frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2 (n_{..k}-1)}} \quad (1)$$

Find one attribute whose PAT value is the l and assume it as X.
Create new node W

if (e = -1) Then
W = N //W will be the root node
else
add W as a child node of N
end if

A^* = (A - best attribute with largest PAT value)
Divide D into left dataset, D1 and right dataset, D2

call **Partial_Association_Tree_Generation** (W, A^*, D, D1, h, 1)

call **Partial_Association_Tree_Generation** (W, A^*, D, D2, h, 0)
}

4.1.2. PRUNE (N)

1. for (each leaf in the Partial_Association_Tree) do
2. if (the paired siblings have the same decision value) then make the parent node as leaf with decision value as label and then remove paired leaves.
   2.1. repeat the process up in the tree until two siblings have different decision values.
   2.2 end if
3. end for

5. Dataset

The dataset named 'Cardiovascular Disease dataset 'is obtained from Kaggle Progression System. There are twelve attributes in the dataset named {chest pain type, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target} of which target is the response variable. The size of this dataset is 70,000. The value of this decision attribute target decides whether the patient has heart problem or not.

6. Results and Discussion

On "Cardiovascular Disease dataset" the proposed algorithm is applied and the dataset is extracted from Kaggle Progression System with 12 attributes and 70000 tuples. A partial association tree with causal relationships is obtained. The node wise division of data along with the correlated attributes with highest partial association test value and the status of the node at each level is given in Table-1
Initially there are 12 attributes including the target attribute in the dataset. One pretest of correlation between each predictor variable versus the decision variable is determined. After this step the size of the dataset becomes 6. For these attributes the proposed algorithm is applied and Partial Association Tree is obtained.

Table.1. Node Wise Distribution of Data

Level of the tree	Attributes after dimensionality reduction	Correlation threshold satisfied attributes	Attribute with highest PAMH value	Number of tuples	Status of the node
0	chest pain type, fbs, restecg, thalach, exang, oldpeak,	chest pain type, fbs, restecg, thalach, exang, oldpeak,	restecg	70000	Partitioned
1	chest pain type, fbs, thalach, exang, oldpeak	chest pain type, fbs, thalach, exang, oldpeak	chest pain type	40737	Partitioned
2	fbs, thalach, exang, oldpeak	fbs, thalach, exang, oldpeak	exang	27866	Partitioned
3	fbs, thalach, oldpeak	fbs, thalach, oldpeak	fbs	18344	Partitioned
4	thalach, oldpeak	Null	Null	5255	No split

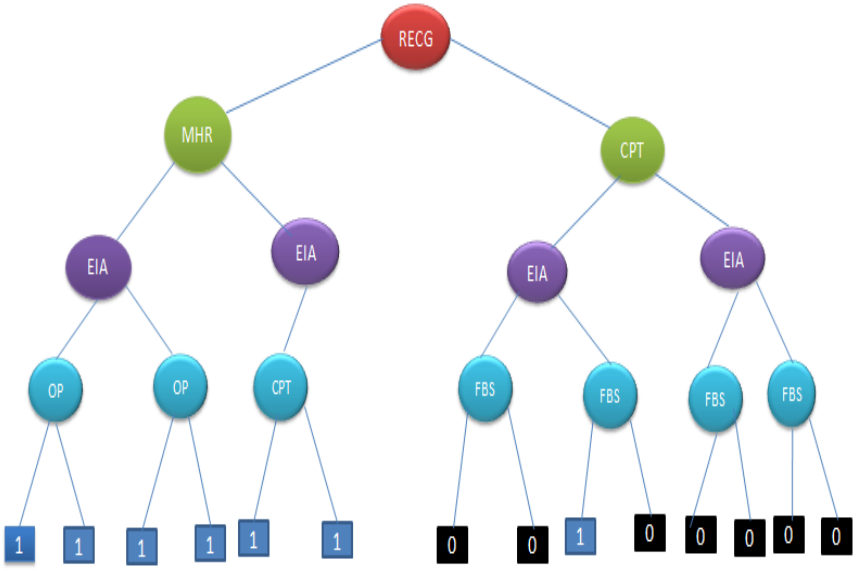

Fig.1. Partial Association Tree without Pruning

*RECG- Resting ECG
*MHR- Maximum Heart Rate
*EIA- Exercise Induced Angina
*OP-Old Peak
* CPT-Chest Pain Type
*FBS- Fasting Blood Sugar

The partial association tree (before pruning) is shown in fig-1. The partial association tree after pruning is represented in fig-2.This tree has limited and less number of paths from root to leaves compared to the number of such paths that can be derived from a normal decision tree.

This big difference is a significant one because this can relieve us from the burden of understanding a large collection of general decision rules derivable from a normal decision tree. The proposed partial association tree model provides portable collection of rules and causal relationships in data that are interpretable and actionable. This type of causal

inference is not possible to achieve through a simple decision tree/classification tree. From the final partial association tree we can derive the causal relationships. The above causal decision tree has three leaves with label 0 (the patient has no heart problem) and two leaves with label 1 (the patient has heart problem).Now we are able to find the causes of various effects/attribute values of the decision variable. In this way it is clear that partial association tree can provide interpretable knowledge through sets of causal relationships. It can provide context specific causal relationships.

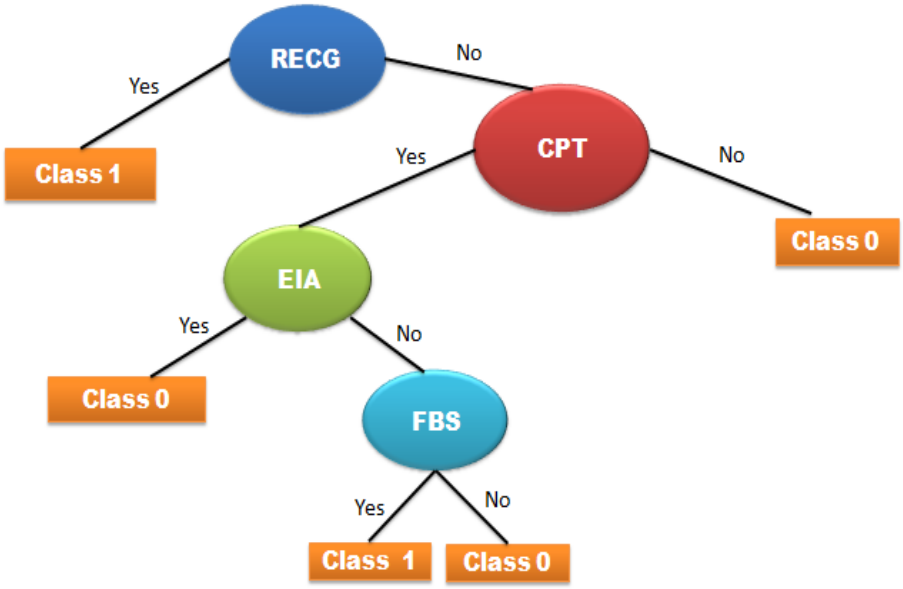

Fig.2. Partial Association Tree (After pruning)

Conclusion

A new method is proposed for finding the casual rules from observational data using Dimensionality Reduction Partial Association Tree. The proposed method is scalable, faster, stable, and efficient than the existing causal relationship discovery method. Many existing method that are based on causal Bayesian network technique are able to find causes consisting of single variable only where as the proposed method able to find causes consisting of combined variables. The proposed method is efficient, effective and an alternative technique for finding causal relationships with respect to the basic original causal Bayesian network method. The method is scalable with respect to large and high dimensional data sets. This type of causal inference is not possible to achieve through a simple decision tree. Compared to a simple decision tree the proposed partial association tree provides portable set of causal rules which are better interpretable. The set of rules based on causal relationships certainly forms an actionable knowledge for better decision making in data analytics.

References

[1]. Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. National Cancer Institute. 22(4) 719–748

[2]. P. R. Rosenbaum, "Design of observational Studies," in The Springer Series, Berlin, Germany: Springer, vol. 1, (2010), pp.65-94.

[3]. P.N. Tan, M. Steinbach and V. Kumar, "Introduction to data mining," in The Pearson, Upper Saddle River, NJ, USA: Pearson, vol. 1(2006) pp.37-69.

[4]. Williamson J, Editor "Bayesian Nets and Causality: Philosophical and Computational Foundation," Oxford University Press,(2005)

[5]. K.J Rothman, S Greenland and T.L. Lash, "Modern epidemiology," in The Library of Congress Catalog, Lippincott Williams and Wilkins, vol.3 (2008) pp.5-31.

[6]. Song JW, Chung KC (2010) Observational Studies: Cohort and Case-Control

[7]. Studies. Plastic and Reconstructive Surgery. 126(6):2234-2242.

[8]. Sreeraman Y and Lakshmana Pandian S (2019), “Data analytics and mining in healthcare with emphasis on causal relationship mining,” Recent Technology and Engineering, 8(4) 195-204.

[9]. Gao G, Bud Mishra, and Daniele Ramazzotti (2018), "Causal Data Science for Financial Stress Testing", Journal of Computational Science. 28(1) 294-304.